# Final Project

Tyler Auger

CS555

*Introduction and Data Description relate to both 1.) + 2.) in assignment outline*

## Introduction

The data in "LifeExpectancyData.csv" gives the life expectancy of world countries and various factors that might impact life expectancy. The original data includes many interesting contributors, such as immunization rates for various diseases, thinness(underweight) percentages of various population age groups, etc; however, in this study a more broad analysis will be performed. The original data consisted of 2938 observations and 22 variables. Below is a link to the original data, which was gathered from kaggle.

Life Expectancy Raw Data

Keeping a general prospective while exploring the data set, a few research questions came to my mind. First, has life expectancy increased from 2000 to 2015? While we know this to be true over the last 100 years, I specifically want to look at the time period relating to the recent tech explosion, i.e., the explosion of the internet and information age (2000 - 2015). Next, I was curious about the impact of some of the variables on life expectancy, and the ability to model life expectancy from these variables. As stated earlier, a very long in depth study could take place to analyze all of the 22 potential explanatory variables, but in this general case I will specifically be looking at the influence from infant deaths(per 1000 population), BMI(average), Total expenditure(general government spending on health as a percentage of total expenditure), and schooling(average number of years attended). Finally, looking at the status of countries (developed vs developing), a two - sample test of proportions can be used to determine if developing countries have a greater risk of having a life expectancy below the mean life expectancy. Further, if a proven disproportion does exist, a logistic regression model can be attempted to determine a quantifiable risk. All together this will allow us to formulate an opinion on the general theme of the study, "What is the impact of the tech and information boom on global life expectancy?"

## Describing the Data

### first 50 observations for idea of data set

*not .html, so can't do paged table (to my understanding), reason for just example - too many observations*

| | Country | Status | Year | Life.expectancy | infant.deaths | BMI | Total.expenditure | Schooling |
|---|---|---|---|---|---|---|---|---|
| 1 | Afghanistan | Developing | 2015 | 65.0 | 62 | 19.1 | 8.16 | 10.1 |
| 2 | Afghanistan | Developing | 2014 | 59.9 | 64 | 18.6 | 8.18 | 10.0 |
| 16 | Afghanistan | Developing | 2000 | 54.8 | 88 | 12.2 | 8.20 | 5.5 |
| 17 | Albania | Developing | 2015 | 77.8 | 0 | 58.0 | 6.00 | 14.2 |
| 18 | Albania | Developing | 2014 | 77.5 | 0 | 57.2 | 5.88 | 14.2 |
| 32 | Albania | Developing | 2000 | 72.6 | 1 | 45.0 | 6.26 | 10.7 |
| 33 | Algeria | Developing | 2015 | 75.6 | 21 | 59.5 | NA | 14.4 |

| | Country | Status | Year | Life.expectancy | infant.deaths | BMI | Total.expenditure | Schooling |
|---|---|---|---|---|---|---|---|---|
| 34 | Algeria | Developing | 2014 | 75.4 | 21 | 58.4 | 7.21 | 14.4 |
| 48 | Algeria | Developing | 2000 | 71.3 | 21 | 44.4 | 3.49 | 10.7 |
| 49 | Angola | Developing | 2015 | 52.4 | 66 | 23.3 | NA | 11.4 |
| 50 | Angola | Developing | 2014 | 51.7 | 67 | 22.7 | 3.31 | 11.4 |
| 64 | Angola | Developing | 2000 | 45.3 | 97 | 15.4 | 2.79 | 4.6 |
| 65 | Antigua and Barbuda | Developing | 2015 | 76.4 | 0 | 47.7 | NA | 13.9 |
| 66 | Antigua and Barbuda | Developing | 2014 | 76.2 | 0 | 47.0 | 5.54 | 13.9 |
| 80 | Antigua and Barbuda | Developing | 2000 | 73.6 | 0 | 38.2 | 4.13 | 0.0 |
| 81 | Argentina | Developing | 2015 | 76.3 | 8 | 62.8 | NA | 17.3 |
| 82 | Argentina | Developing | 2014 | 76.2 | 8 | 62.2 | 4.79 | 17.3 |
| 96 | Argentina | Developing | 2000 | 74.1 | 12 | 54.0 | 9.21 | 15.0 |
| 97 | Armenia | Developing | 2015 | 74.8 | 1 | 54.9 | NA | 12.7 |
| 98 | Armenia | Developing | 2014 | 74.6 | 1 | 54.1 | 4.48 | 12.7 |
| 112 | Armenia | Developing | 2000 | 72.0 | 1 | 47.1 | 6.25 | 11.2 |
| 113 | Australia | Developed | 2015 | 82.8 | 1 | 66.6 | NA | 20.4 |
| 114 | Australia | Developed | 2014 | 82.7 | 1 | 66.1 | 9.42 | 20.4 |
| 128 | Australia | Developed | 2000 | 79.5 | 1 | 58.2 | 8.80 | 20.4 |
| 129 | Austria | Developed | 2015 | 81.5 | 0 | 57.6 | NA | 15.9 |
| 130 | Austria | Developed | 2014 | 81.4 | 0 | 57.1 | 11.21 | 15.9 |
| 144 | Austria | Developed | 2000 | 78.1 | 0 | 5.1 | 1.60 | 15.4 |
| 145 | Azerbaijan | Developing | 2015 | 72.7 | 5 | 52.5 | NA | 12.7 |
| 146 | Azerbaijan | Developing | 2014 | 72.5 | 5 | 51.5 | 6.40 | 12.2 |
| 160 | Azerbaijan | Developing | 2000 | 66.6 | 9 | 42.1 | 4.67 | 10.1 |
| 161 | Bahamas | Developing | 2015 | 76.1 | 0 | 64.5 | NA | 12.6 |
| 162 | Bahamas | Developing | 2014 | 75.4 | 0 | 63.8 | 7.74 | 12.6 |
| 176 | Bahamas | Developing | 2000 | 72.6 | 0 | 54.4 | 5.21 | 12.0 |
| 177 | Bahrain | Developing | 2015 | 76.9 | 0 | 63.6 | NA | 14.5 |
| 178 | Bahrain | Developing | 2014 | 76.8 | 0 | 62.9 | 4.98 | 14.5 |
| 192 | Bahrain | Developing | 2000 | 74.5 | 0 | 54.5 | 3.51 | 13.2 |
| 193 | Bangladesh | Developing | 2015 | 71.8 | 92 | 18.3 | NA | 10.2 |
| 194 | Bangladesh | Developing | 2014 | 71.4 | 98 | 17.7 | 2.82 | 10.0 |
| 208 | Bangladesh | Developing | 2000 | 65.3 | 231 | 1.4 | 2.33 | 7.3 |
| 209 | Barbados | Developing | 2015 | 75.5 | 0 | 54.5 | NA | 15.3 |
| 210 | Barbados | Developing | 2014 | 75.4 | 0 | 53.7 | 7.47 | 15.3 |
| 224 | Barbados | Developing | 2000 | 73.3 | 0 | 43.0 | 5.16 | 14.0 |
| 225 | Belarus | Developing | 2015 | 72.3 | 0 | 62.3 | NA | 15.6 |
| 226 | Belarus | Developing | 2014 | 72.0 | 0 | 61.7 | 5.69 | 15.7 |
| 240 | Belarus | Developing | 2000 | 68.0 | 1 | 54.4 | 6.13 | 13.1 |
| 241 | Belgium | Developed | 2015 | 81.1 | 0 | 63.7 | NA | 16.6 |
| 242 | Belgium | Developed | 2014 | 89.0 | 0 | 63.4 | 1.59 | 16.3 |
| 256 | Belgium | Developed | 2000 | 77.6 | 1 | 58.1 | 8.12 | 18.0 |
| 257 | Belize | Developing | 2015 | 71.0 | 0 | 5.9 | NA | 12.8 |
| 258 | Belize | Developing | 2014 | 70.0 | 0 | 5.1 | 5.79 | 12.8 |

After importing the raw data into Rstudio, I then had to subset the data, containing only the columns and observations I needed for the study. In this case, I chose to restrict the observations to those from 2000, 2014, and 2015. This was needed to explore the possible significant change in life expectancy between the start of the tech and information explosion, to the latest available time from the data (since we are still very much in this explosion), i.e., data from 2000 and 2015. I then restricted the columns to only the variables needed for

my broad modeling mentioned in the introduction, and the year 2014 (since not enough data was available in regards to total expenditure for the year 2015). I also decided to keep Country as a reference if wanting to compare my data set to the raw data; however, it is not needed to answer the proposed research questions. The final, clean data set for this study contained 549 observations and 8 variables, with 183 observations per year. Also of importance when modeling the data, some of the variables have "NA" observations. Thus the rows containing "NA" will be dropped from the data. Further, to insure that the data is independent and random, only the data relating to the most recent year with enough data, 2014, will be used when modeling.

## Methods

*assignment outline 3.)*

First, it is important to note that all significance tests will be tested at the alpha level of 0.05.

A two-sample mean test can be used to determine if there was a significant change in the mean life expectancy between 2000 and 2015. First a boxplot can be used to visually explore the difference in life expectancy distributions between 2000 and 2015.

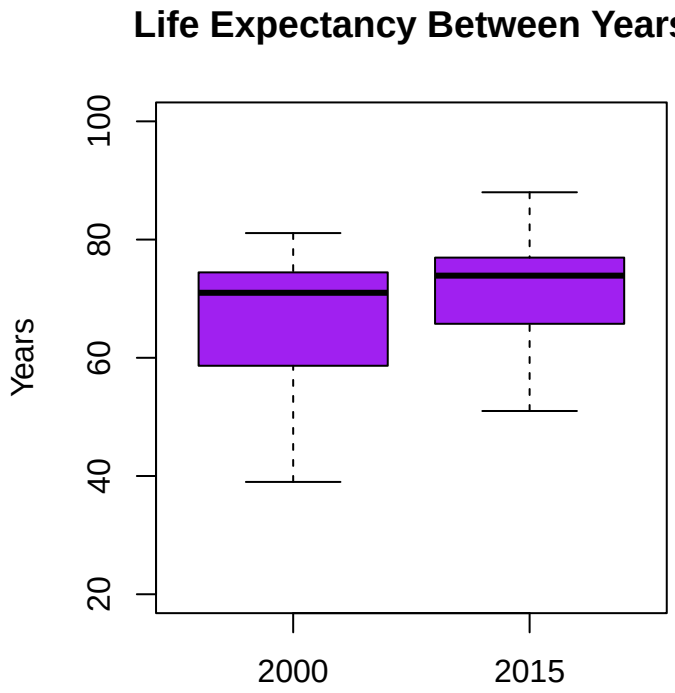
When modeling the data, a multiple linear regression may be used to test if the variables infant deaths, BMI, Total expenditure, and schooling significantly model life expectancy. A scatterplot matrix will also help us look at correlations among the variables, and address any issues, i.e. collinearity, normality, etc. If the general model is deemed significant, we can then test the significance of each individual variable, and measure the fit of an overall final model.

Finally, looking specifically at the status (developed vs developing) of each observation in 2014 and whether the life expectancy relating to a country's status is above the mean life expectancy from 2014, we can formally test whether the proportion of developed and developing countries with life expectancy greater than the mean life expectancy of 2014 are equal. If there is a difference, we can then perform logistic regression. If proven significant, we can develop an odds ratio on the possible increased risk of having an above average life expectancy based on country status.

All together, this will help us formulate an informed opinion to the impact of information accessibility (the symbiotic relationship between tech and information availability - such as the global availability of a "smart phone" and internet) on life expectancy.

## Results: Life Expectancy 2000 vs 2015

*assignment outline 4.) split between each "Results:"*




**summary of life expectancy for 2000**

```
##  Life.expectancy
##  Min.   :39.00
##  1st Qu.:58.65
##  Median :71.00
##  Mean   :66.75
##  3rd Qu.:74.45
##  Max.   :81.10
```

```
## [1] "With a standard deviation of 10.2955280717692"
```

**summary of life expectancy for 2015**

```
##  Life.expectancy
##  Min.   :51.00
##  1st Qu.:65.75
##  Median :73.90
##  Mean   :71.62
##  3rd Qu.:76.95
##  Max.   :88.00
```

```
## [1] "With a standard deviation of 8.1237061476453"
```

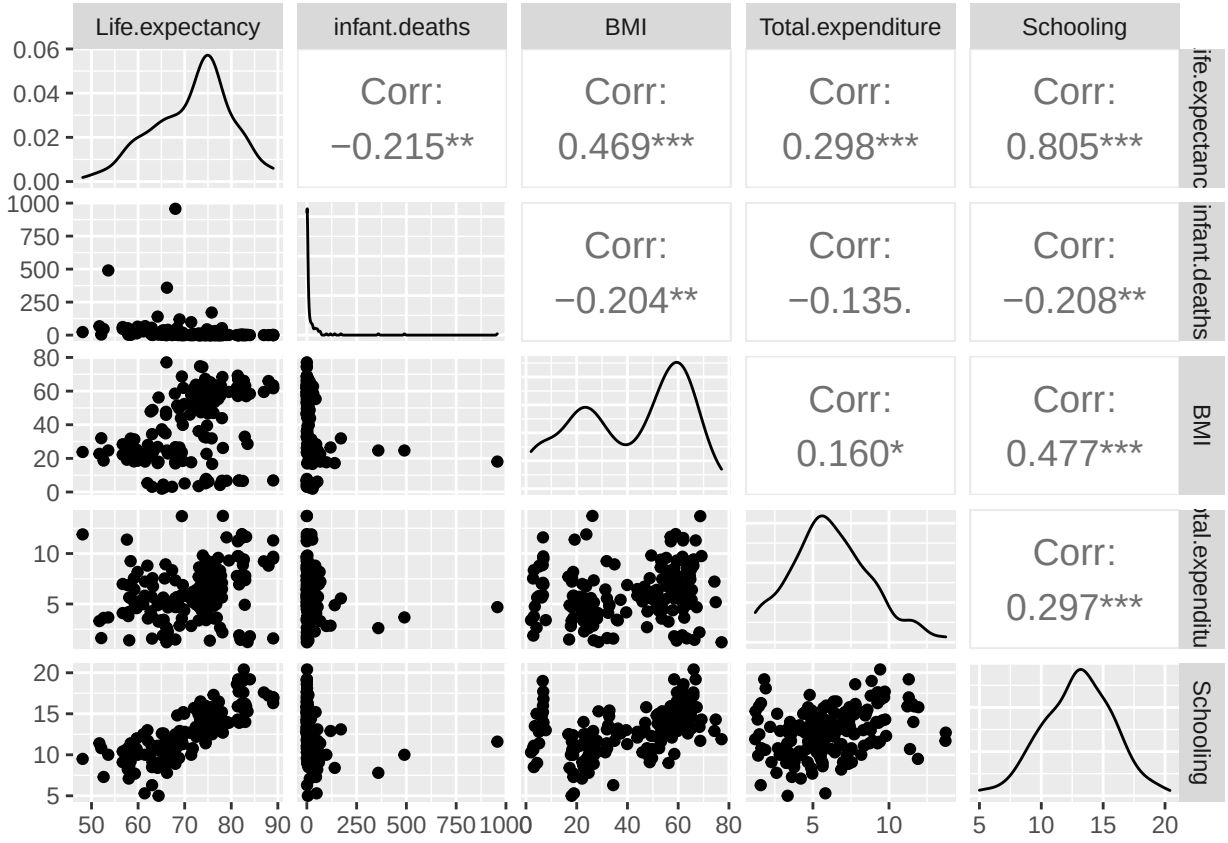
Looking at the boxplots and summary of the two samples, there seems to definitely be a difference in the mean life expectancy between the 15 years. Also, the standard deviation of life expectancy decreased from

2000 to 2015. Further, it looks like there are no outlier problems. While this insinuates that there has been a significant increase in life expectancy, a two-sample test of means will confirm suspicions.

```
##
##  Welch Two Sample t-test
##
## data:  data_LE2000 and data_LE2015
## t = -5.02, df = 345.32, p-value = 8.295e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.773453 -2.959880
## sample estimates:
## mean of x mean of y
##  66.75027  71.61694
```

From the results above, there is significant evidence at the alpha level of 0.05 that the mean life expectancy is different between 2000 and 2015. Further, looking at the difference in sample means and the confidence interval provided above, with 95% confidence the true mean difference in life expectancy between 2000 and 2015 is -6.773453 and -2.959880. This confirms that life expectancy has significantly increased during the tech and information boom between 2000 and 2015.

## Results: Modeling Life Expectancy




Looking at the scatterplots to graphically analyze and understand the correlations among the variables, there seems to be a massive correlation between schooling and life expectancy. While this is not unexpected, it

is surprising to see such a strong correlation, especially compared to the other variables. Further, it looks like there may be an issue using infant deaths in the model. A negative correlation does not make much sense, since young person deaths would drastically lower the mean life expectancy, but this will be explored further after globally testing the strength of the model, or re-trying the global test without infant deaths if the model is not deemed significant. Also, there seems like there could be an issue with normality for both infant deaths and BMI; however, based on the correlations, these may be ineffective in the model anyways after further analysis. Finally, collinearity does not seem to be an issue between the explanatory variables.

```
##
## Call:
## lm(formula = data_mlr$Life.expectancy ~ data_mlr$infant.deaths +
##     data_mlr$BMI + data_mlr$Total.expenditure + data_mlr$Schooling)
##
## Coefficients:
##                 (Intercept)      data_mlr$infant.deaths
##                   40.946582                   -0.003167
##                data_mlr$BMI  data_mlr$Total.expenditure
##                    0.041150                    0.192494
##          data_mlr$Schooling
##                    2.160707

## Analysis of Variance Table
##
## Response: data_mlr$Life.expectancy
##                              Df Sum Sq Mean Sq F value    Pr(>F)
## data_mlr$infant.deaths        1  549.3   549.3  22.618 4.266e-06 ***
## data_mlr$BMI                  1 2252.7  2252.7  92.756 < 2.2e-16 ***
## data_mlr$Total.expenditure    1  547.9   547.9  22.558 4.385e-06 ***
## data_mlr$Schooling            1 4538.1  4538.1 186.858 < 2.2e-16 ***
## Residuals                   166 4031.6    24.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] "Critical F(df = 4, 166, alpha = 0.05) = 2.42610565724194"

## [1] "R^2 = 0.661770263995348"

## [1] "F(df = 3, 98) = 81.1976684256679"
```

The least squares regression equation that predicts life expectancy(y) from infant deaths(X1), BMI(X2), Total expenditure(X3), and Schooling(X4) is:

y(hat) = 40.946582 - 0.003167 * X(1) + 0.041150 * X(2) + 0.192494 * X(3) + 2.160707 * X(4)

And based on the information above, there is significant evidence at the alpha level of 0.05 that there is at least one slope coefficient that is different than 0. Thus, there is significant evidence of linear association between the response and at least one of the explanatory variables. Looking at the R^2 value, the model seems to do a decent job of explaining the variation of life expectancy. However, as mentioned earlier, each variable needs to be looked at to determine their significance in the overall model.

```
##
## Call:
## lm(formula = data_mlr$Life.expectancy ~ data_mlr$infant.deaths +
```

```
##     data_mlr$BMI + data_mlr$Total.expenditure + data_mlr$Schooling)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.570  -2.948   0.417   3.360  11.282
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 40.946582   1.883641  21.738   <2e-16 ***
## data_mlr$infant.deaths      -0.003167   0.004424  -0.716   0.4752
## data_mlr$BMI                 0.041150   0.020407   2.016   0.0454 *
## data_mlr$Total.expenditure   0.192494   0.150814   1.276   0.2036
## data_mlr$Schooling           2.160707   0.158067  13.670   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.928 on 166 degrees of freedom
## Multiple R-squared:  0.6618, Adjusted R-squared:  0.6536
## F-statistic:  81.2 on 4 and 166 DF,  p-value: < 2.2e-16
```

The summary of the model above clearly shows that infant deaths and Total expenditure are insignificant to the model when adjusting for the other explanatory variables, with very large p-values. Since infant deaths seems to be the least significant to the model, that will be removed first to see how the model is affected.

**Without infant deaths**

```
##
## Call:
## lm(formula = data_mlr$Life.expectancy ~ data_mlr$BMI + data_mlr$Total.expenditure +
##     data_mlr$Schooling)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.5456  -2.9564   0.3747   3.3502  11.4982
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 40.59566    1.81606  22.354   <2e-16 ***
## data_mlr$BMI                 0.04292    0.02023   2.122   0.0353 *
## data_mlr$Total.expenditure   0.20070    0.15016   1.337   0.1832
## data_mlr$Schooling           2.17265    0.15695  13.843   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.921 on 167 degrees of freedom
## Multiple R-squared:  0.6607, Adjusted R-squared:  0.6546
## F-statistic: 108.4 on 3 and 167 DF,  p-value: < 2.2e-16
```

With barely any change to the R^2 value, this seems like a better model. Since Total expenditure is still insignificant when adjusting for the other explanatory variables, that will also be removed, most likely producing the best predictive model.

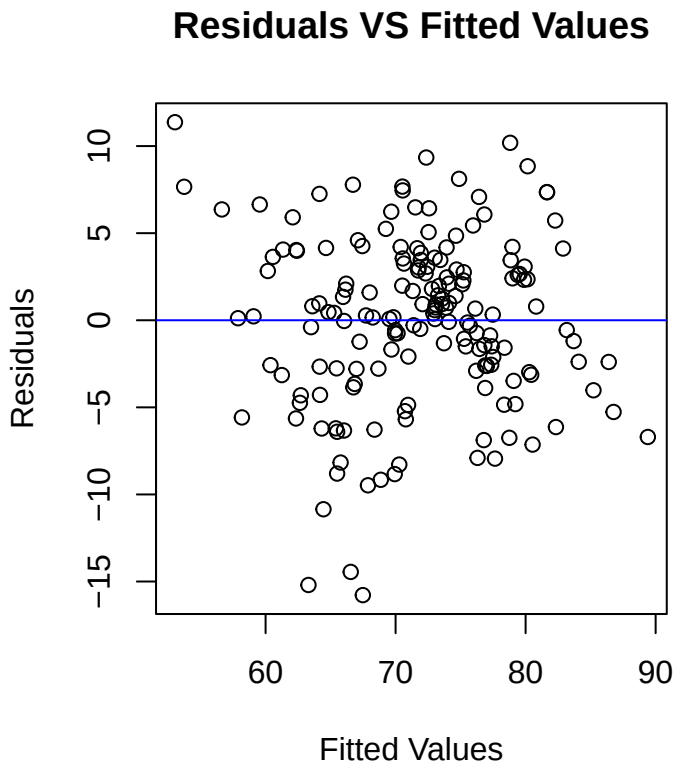
**Without infant deaths and Total Expenditure**

```
##
```

```
## Call:
## lm(formula = data_mlr$Life.expectancy ~ data_mlr$BMI + data_mlr$Schooling)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7799  -2.7845   0.3906   3.3508  11.3704
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        41.11694    1.77783  23.128   <2e-16 ***
## data_mlr$BMI        0.04351    0.02027   2.147   0.0333 *
## data_mlr$Schooling  2.22589    0.15217  14.628   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.932 on 168 degrees of freedom
## Multiple R-squared:  0.6571, Adjusted R-squared:  0.653
## F-statistic:   161 on 2 and 168 DF,  p-value: < 2.2e-16
```

Following the same test of significance as before and seeing the very small reduction in R^2, as expected the model that best predicts life expectancy(y) from the available explanatory variables is:

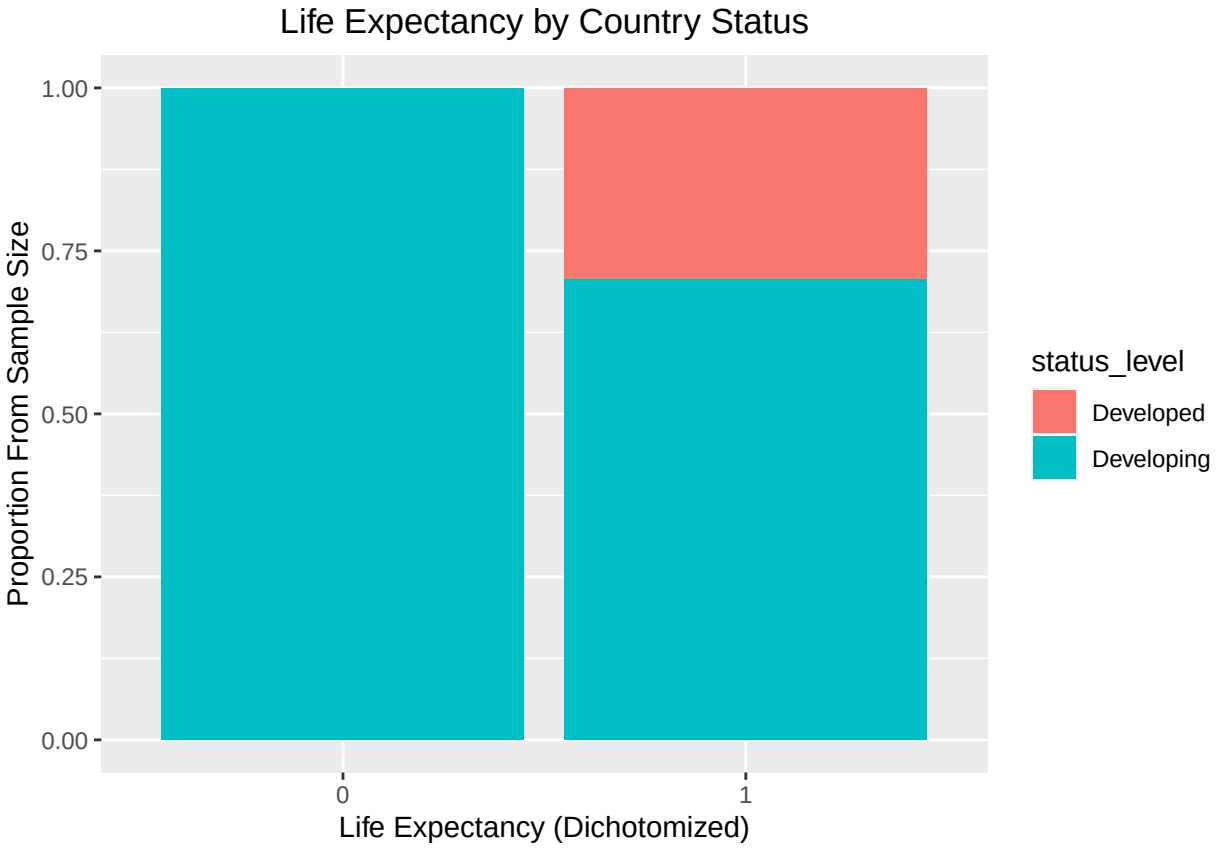
y = 0.04351 * X1(BMI) + 2.22589 * X2(Schooling) + 41.11694 + e




The fit of the model also seems reasonable - the points seem random, with variation for the most part constant, and linearity and normality also holding.

## Results: Test of Proportions and Logistic Regression

**Life Expectancy 1 (Successes) >= 71.7655; < 71.7655 is 0 (Failure)**




| Population | Description | Sample.Size | Successes | Failures | Sample.Proportions |
|---|---|---|---|---|---|
| 1 | Developed | 29 | 29 | 0 | 1.0000000 |
| 2 | Developing | 142 | 70 | 72 | 0.4929577 |

Looking at the table and graph above, a visual, quick analysis of the sample data seems to insinuate that a greater proportion of developed countries compared to developing countries have a life expectancy above the mean life expectancy of all countries in 2014; however, further analysis from a two-sample test for proportions will provide better insight.

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(29, 70) out of c(29, 142)
## X-squared = 25.398, df = 1, p-value = 4.664e-07
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.4248121 0.5892724
## sample estimates:
##    prop 1    prop 2
## 1.0000000 0.4929577
```

```
## [1] "The estimate of the risk difference is 0.507042253521127"
```

From the above data, it is clear there is significant evidence at the alpha level of 0.05 that the proportion of developed countries have a higher life expectancy than developing countries. Since we have significant evidence of proportional differences, a logistic regression, if proven significant, might give insight to the quantifiable advantage developed countries might have over developing countries regarding life expectancy.

```
##
## Call:
## glm(formula = data_group$LE_level ~ data_group$status_level,
##     family = binomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.16547  -1.16547   0.00013   1.18940   1.18940
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  37.16    2422.45   0.015    0.988
## data_group$status_level     -18.59    1211.22  -0.015    0.988
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 232.78  on 170  degrees of freedom
## Residual deviance: 196.83  on 169  degrees of freedom
## AIC: 200.83
##
## Number of Fisher Scoring iterations: 17
```

From the output above testing the significance of the logistic regression model, we fail to reject the null hypothesis, and that there is evidence that at the alpha level of 0.05, $B(1) = 0$. Therefore, we have proven that the developed countries have a greater proportion of life expectancy above the world mean, but there is not significant evidence of an association between Country status and the risk of having a lower/higher life expectancy. Relating this to the other statistical tests done earlier, this gives a very interesting, philosophical insight into the possible biggest influence of life expectancy.

## Conclusion

Information is power.

From the above studies, it is clear that the more information an individual, or country, might be able to access, the greater the life expectancy. This is insinuated by looking at the results provided when answering the three proposed questions outlining the research theme. First, showing that life expectancy significantly increased between 2000 and 2015, the initial hypothesis that life expectancy increased during the tech and information boom can be more deeply explored. Next, after attempting to form a predictive model that fit the distribution of life expectancy from the sample provided; surprisingly, schooling was by far the most influential contributor to life expectancy. This, along with the fact that life expectancy significantly increased over the tech and information boom, starts to show a consistent pattern for life expectancy and knowledge, and in this setting, deemed as information accessibility. Finally, while the proportion of developed countries having a higher than average life expectancy compared to developing countries was confirmed, the notion of developed countries having a measurable advantage over developing countries in regards to life expectancy was disproven. While there is a puzzling nature to this, a possibility does comes to mind when relating this to the importance of schooling in the formulated model fitting life expectancy to amount of schooling

and BMI. Since developed vs developing countries normally relates to infrastructure and GDP, it does not necessarily relate to the access of information and education through personal technology - remote places can now have access to phones and online-schooling, internet is effectively available everywhere, and collaboration and the ability to share information and intelligence on a global scale has never been greater. While not a definitive direct answer to the research question posed in the introduction, a strong, educated insight has been developed:

**The importance of providing accessible, real education and information to people on an individual level is paramount to the lengthening of human life on a global scale**

While some strong convictions can be formulated from this study, it does have some potential pitfalls. First, some factors that may completely change the influence of schooling on life expectancy were not included in the data set. For example, child starvation was available in the raw data, but left out of the clean data. Also clean water accessibility was not even included in the raw data. Another pitfall were some of the inferences made from looking at all three studies together, such as the fact that there is no direct meaning of what contributes to being a "developed" country. In future studies, it would be very beneficial to define what makes a country "developed" in the data. Also, deeper insights could be developed by expanding on the variable Schooling. Instead, it could be a grouped factor variable relating to a broader category of information accessibility. One group could very well be years of schooling,and others could consist of smart phone use per capita, internet access, etc.

In conclusion, information is power, and individually this power can be fuel for human life.